jon

Network Working Group
Request for Comments: 1077

Gigabit Working Group
B. Leiner, Editor
November 1988

Critical Issues in High Bandwidth Networking

Status of this Memo

This memo presents the results of a working group on High Bandwidth Networking. This RFC is for your information and you are encouraged to comment on the issues presented. Distribution of this memo is unlimited.

ABSTRACT

At the request of Maj. Mark Pullen and Maj. Brian Boesch of DARPA, an ad-hoc working group was assembled to develop a set of recommendations on the research required to achieve a ubiquitous high-bandwidth network as discussed in the FCCSET recommendations for Phase III.

This report outlines a set of research topics aimed at providing the technology base for an interconnected set of networks that can provide highbandwidth capabilities. The suggested research focus draws upon ongoing research and augments it with basic and applied components. The major activities are the development and demonstration of a gigabit backbone network, the development and demonstration of an interconnected set of networks with gigabit throughput and appropriate management techniques, and the development and demonstration of the required overall architecture that allows users to gain access to such high bandwidth.

1. Introduction and Summary

1.1. Background

The computer communications world is evolving toward both high-bandwidth capability and high-bandwidth requirements. The recent workshop conducted under the auspices of the FCCSET Committee on High Performance Computing [1] identified a number of areas where extremely high-bandwidth networking is required to support the scientific research community. These areas range from remote graphical visualization of supercomputer results through the movement of high rate sensor data from space to the ground-based scientific investigator. Similar requirements exist for other applications, such as military command and control (C2) where there is a need to quickly access and act on data obtained from real-time sensors. The workshop identified requirements for switched high-bandwidth service in excess of 300 Mbit/s to a single user, and the need to support service in the range of a Mbit/s on a low-duty-cycle basis to millions of researchers. When added to the needs of the military and commercial users, the aggregate requirement for communications service adds up to many billions of bits per second. The results of this workshop were incorporated into a report by the FCCSET [2].

Fortunately, technology is also moving rapidly. Even today, the installed base of fiber optics communications allows us to consider aggregate bandwidths in the range of Gbit/s and beyond to limited geographical regions. Estimates arrived at in the workshop lead one to believe that there will be available raw bandwidth approaching terabits per second.

The critical question to be addressed is how this raw bandwidth can be used to satisfy the requirements identified in the workshop: 1) provide bandwidth on the order of several Gbit/s to individual users, and 2) provide modest bandwidth on the order of several Mbit/s to a large number of users in a cost-effective manner through the aggregation of their traffic.

Through its research funding, the Defense Advanced Research Projects Agency (DARPA) has played a central role in the development of packet-oriented communications, which has been of tremendous benefit to the U.S. military in terms of survivability and interoperability. DARPA-funded research has resulted in the ARPANET, the first packet-switched network; the SATNET, MATNET and Wideband Network, which demonstrated the efficient utilization of shared-access satellite channels for communications between geographically diverse sites;

packet radio networks for mobile tactical environments; the Internet and TCP/IP protocols for interconnection and interoperability between heterogeneous networks and computer systems; the development of electronic mail; and many advances in the areas of network security, privacy, authentication and access control for distributed computing environments. Recognizing DARPA's past accomplishments and its desire to continue to take a leading role in addressing these issues, this document provides a recommendation for research topics in gigabit networking. It is meant to be an organized compendium of the critical research issues to be addressed in developing the technology base needed for such a high bandwidth ubiquitous network.

1.2. Ongoing Activities

The OSTP report referred to above recommended a three-phase approach to achieving the required high-bandwidth networking for the scientific and research community. Some of this work is now well underway. An ad-hoc committee, the Federal Research Internet Coordinating Committee (FRICC) is coordinating the interconnection of the current wide area networking systems in the government; notably those of DARPA, Department of Energy (DoE), National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), and the Department of Health and Human Services (HHS). In accordance with Phases I and II of the OSTP report, this activity will provide for an interconnected set of networks to support research and other scholarly pursuits, and provide a basis for future networking for this community. The networking is being upgraded through shared increased bandwidth (current plans are to share a 45 Mbit/s backbone) and coordinated interconnection with the rest of the world. In particular, the FRICC is working with the European networking community under the auspices of another ad-hoc group, the Coordinating Committee for Intercontinental Research Networks (CCIRN), to establish effective US-Europe networking.

However, as the OSTP recommendations note, the required bandwidth for the future is well beyond currently planned public, private, and government networks. Achieving the required gigabit networking capabilities will require a strong research activity. There is considerable ongoing research in relevant areas that can be drawn upon; particularly in the areas of high-bandwidth communication links, high-speed computer switching, and high-bandwidth local area networks. Appendix A provides some pointers to current research efforts.

1.3. Document Overview

This report outlines a set of research topics aimed at providing the technology base for an interconnected set of networks that can provide the required high-bandwidth capabilities discussed above. The suggested research focus draws upon ongoing research and augments it with basic and applied components. The major activities are the development and demonstration of a Gigabit Backbone network (GB) [3], the development and demonstration of an interconnected set of networks with gigabit throughput and appropriate management techniques, and the development and demonstration of the required overall architecture that allows users to gain access to such high bandwidth. Section 2 discusses functional and performance goals along with the anticipated benefits to the ultimate users of such a system. Section 3 provides the discussion of the critical research issues needed to achieve these goals. It is organized into the major areas of technology that need to be addressed: general architectural issues, high-bandwidth switching, high-bandwidth host interfaces, network management algorithms, and network services. The discussion in some cases contains examples of ongoing relevant research or potential approaches. These examples are intended to clarify the issues and not to propose that particular approach. A discussion of the relationship of the suggested research to other ongoing activities and optimal methods for pursuing this research is provided in Section 4.

2. Functional and Performance Goals

In this section, we provide an assessment of the types of services a GN (four or five orders of magnitude faster than the current networks) should provide to its users. In instances where we felt there would be a significant impact on performance, we have provided an estimate of the amount of bandwidth needed and delay allowable to provide these services.

2.1. Networking Application Support

It is envisioned that the GN will be capable of supporting all of the following types of networking applications.

Currently Provided Packet Services

It is important that the network provide the users with the equivalent of services that are already available in packet-switched networks, such as interactive data exchange, mail service, file transfer, on-line access to remote computing resources, etc., and allow them to expand to other more advanced services to meet their needs as they become available.

Multi-Media Mail

This capability will allow users to take advantage of different media types (e.g., graphics, images, voice, and video as well as text and computer data) in the transfer of messages, thereby increasing the effectiveness of message exchange.

Multi-Media Conferencing

Such conferencing requires the exchange of large amounts of information in short periods of time. Hence the requirement for high bandwidth at low delay. We estimate that the bandwidth would range from 1.5 to 100 Mbit/s, with an end-to-end delay of no more than a few hundred msec.

Computer-Generated Real-time Graphics

Visualizing computer results in the modern world of supercomputers requires large amounts of real time graphics. This in turn will require about 1.5 Mbit/s of bandwidth and no more than several hundred msec. delay.

High-Speed Transaction Processing

One of the most important reasons for having an ultra-high-speed network is to take advantage of supercomputing capability. There are several scenarios in which this capability could be utilized. For example, there could be instances where a non-supercomputer may require a supercomputer to perform some processing and provide some intermediate results that will be used to perform still further processing, or the exchange may be between several supercomputers operating in tandem and periodically exchanging results, such as in a battle management, war gaming, or process control applications. In such cases, extremely short response times are necessary to accomplish as many as hundreds of interactions in real time. This requires very high bandwidth, on the order of 100 Mbit/s, and minimum delay, on the order of hundreds of msec.

Wide-Area Distributed Data/Knowledge Base Management Systems

Computer-stored data, information, and knowledge is distributed around the country for a variety of reasons. The ability to perform complex queries, updates, and report generation as though many large databases are one system would be extremely powerful, yet requires low-delay, high-bandwidth communication for interactive use. The Corporation for National Research Initiatives (NRI) has promoted the notion of a National Knowledge base with these characteristics. In particular, an attractive approach is to cache views at the user sites, or close by to allow efficient repeated queries and multi-relation processing for relations on different nodes. However, with caching, a processing activity may incur a miss in the midst of a query or update, causing it to be delayed by the time required to retrieve the missing relation or portion of relation. To minimize the overhead for cache directories, both at the server and client sites, the unit of caching should be large---say a megabyte or more. In addition, to maintain consistency at the caching client sites, server sites need to multicast invalidations and/or updates. Communication requirements are further increased by replication of the data. The critical parameter is latency for cache misses and consistency operations. Taking the distance between sites to be on average 1/4 the diameter of the country, a one Gbit/s data rate is required to reduce the transmission time to be roughly the same as the propagation delay, namely around 8 milliseconds for this size of unit. Note that this application is supporting far more sophisticated queries and updates than normally associated with transaction processing, thus requiring larger amount of data to be transferred.

2.2. Types of Traffic and Communications Modes

Different types of traffic may impose different constraints in terms of throughput, delay, delay dispersion, reliability and sequenced delivery. Table 1 summarizes some of the main characteristics of several different types of traffic.

Table 1: Communication Traffic Requirements

Traffic Type	Delay Requirement	Throughput Requirement	Error-free Sequenced Delivery
Interactive Simulation	Low	Moderate-High	No
Network Monitoring	Moderate	Low	No
Virtual Terminal	Low	Low	Yes
Bulk Transfer	High	High	Yes
Message	Moderate	Moderate	Yes
Voice	Low, constant	Moderate	No
Video	Low, constant	High	No
Facsimile	Moderate	High	No
Image Transfer	Variable	High	No
Distributed Computing	Low	Variable	Yes
Network Control	Moderate	Low	Yes

The topology among users can be of three types: point-to-point (one-to-one connectivity), multicast (one sender and multiple receivers), and conferencing (multiple senders and multiple receivers). There are three types of transfers that can take place among users. They are connection-oriented network service, connectionless network service, and stream or synchronous traffic. Connection and connectionless services are asynchronous. A connection-oriented service assumes and provides for relationships among the multiple packets sent over the connection (e.g., to a common destination) while connectionless service assumes each packet is a complete and separate entity unto itself. For stream or synchronous service a reservation scheme is used to set up and guarantee a constant and steady amount of bandwidth between any two subscribers.

2.3. Network Backbone

The GB needs to be of high bandwidth to support a large population of users, and additionally to provide high-speed connectivity among certain subscribers who may need such capability (e.g., between two supercomputers). These users may access the GN from local area networks (LANs) directly connected to the backbone or via high-speed intermediate regional networks. The backbone must also minimize end-to-end delay to support highly interactive high-speed (supercomputer) activities.

It is important that the LANs that will be connected to the GN be permitted data rates independent of the data rates of the GB. LAN speeds should be allowed to change without affecting the GB, and the GB speeds should be allowed to change without affecting the LANs. In this way, development of the technology for LANs and the GB can proceed independently.

Access rate requirements to the GB and the GN will vary depending on user requirements and local environments. The users may require access rates ranging from multi-kbit/s in the case of terminals or personal computers connected by modems up to multi-Mbit/s and beyond for powerful workstations up to the Gbit/s range for high-speed computing and data resources.

2.4. Directory Services

Directory services similar to those found in CCITT X.500/ISO DIS 9594 need to be provided. These include mapping user names to electronic mail addresses, distribution lists, support for authorization checking, access control, and public key encryption schemes, multimedia mail capabilities, and the ability to keep track of mobile users (those who move from place to place and host computer to host computer). The directory services may also list facilities available to users via the network. Some examples are databases, supercomputing or other special-purpose applications, and on-line help or telephone hotlines.

The services provided by X.500 may require some extension for GN. For example, there is no provision for multilevel security, and the approach taken to authentication must be studied to ensure that it meets the requirements of GN and its user community.

2.5. Network Management and Routing

The objective of network management is to ensure that the network functions smoothly and efficiently, and consists of the following: accounting, security, performance monitoring, fault isolation and configuration control.

Accounting ensures that users are properly billed for the services that the network provides. Accounting enforces a tariff; a tariff expresses a usage policy. The network need only keep track of those items addressed by the tariff, such as allocated bandwidth, number of packets sent, number of ports used, etc. Another type of accounting may need to be supported by the network to support resource sharing, namely accounting analogous to telephone "900" numbers. This accounting performed by the network on behalf of resource providers and consumers is a pragmatic solution to the problem of getting the users and consumers into a financial relationship with each other which has stymied previous attempts to achieve widespread use of specialized resources.

Performance monitoring is needed so that the managers can tell how the network is performing and take the necessary actions to keep its performance at a level that will provide users with satisfactory service. Fault isolation using technical control mechanisms is needed for network maintenance. Configuration management allows the network to function efficiently.

Several new types of routing will be required by GN. In addition to true type-of-service, needed to support diverse distributed applications, real-time applications, interactive applications, and bulk data transfer, there will be need for traffic controls to enforce various routing policies. For example, policy may dictate that traffic from certain users, applications, or hosts may not be permitted to traverse certain segments of the network. Alternatively, traffic controls may be used to promote fairness; that is, to make sure that busy link or network segment isn't dominated by a particular source or destination. The ability of applications to reserve network bandwidth in advance of its use, and the use of strategies such as soft connections, will also require development of new routing algorithms.

2.6. Network Security Requirements

Security is a critical factor within the GN and one of those features that are difficult to provide. It is envisioned that both

unclassified and classified traffic will utilize the GN, so protection mechanisms must be an integral part of the network access strategy. Features such as authentication, integrity, confidentiality, access control, and nonrepudiation are essential to provide trusted and secure communication services for network users.

A subscriber must have assurance that the person or system he is exchanging information with is indeed who he says he is. Authentication provides this assurance by verifying that the claimed source of a query request, control command, response, etc., is the actual source. Integrity assures that the subscriber's information (such as requests, commands, data, responses, etc.) is not changed, intentionally or unintentionally, while in transit or by replays of earlier traffic. Unauthorized users (e.g., intruders or network viruses) would be denied use of GN assets through access control mechanisms which verify that the authenticated source is authorized to receive the requested information or to initiate the specified command. In addition, nonrepudiation services can be offered to assure a third party that the transmitted information has not been altered. And finally, confidentiality will ensure that the contents of a message are not divulged to unauthorized individuals. Subscribers can decide, based upon their own security needs and particular activities, which of these services are necessary at a given time.

3. Critical Research Issues

In the section above, we discussed the goals of a research program in gigabit networking; namely to provide the technology base for a network that will allow gigabit service to be provided in an effective way. In this section, we discuss those issues which we feel are critical to address in a research program to achieve such goals.

3.1. General Architectural Issues

In the last generation of networks, it was assumed that bandwidth was the scarce resource and the design of the switch was dictated by the need to manage and allocate the bandwidth effectively. The most basic change in the next generation network is that the speeds of the trunks are rising faster than the speeds of the switching elements.

This change in the balance of speeds has manifested itself in several ways. In most current designs for local area networks, where

bandwidth is not expensive, the design decision was to trade off effective use of the bandwidth for a simplified switching technique. In particular, networks such as Ethernet use broadcast as the normal distribution method, which essentially eliminates the need for a switching element.

As we look at still higher speed networks, and in particular networks in which the bandwidth is still the expensive component, we must design new options for switching which will permit effective use of bandwidth without the switch itself becoming the bottleneck.

The central thrust of new research must thus be to explore new network architectures that are consistent with these very different speed assumptions.

The development of computer communications has been tremendously distorted by the characteristics of wide-area networking: normally high cost, low speed, high error rate, large delay. The time is ripe for a revolution in thinking, technology, and approaches, analogous to the revolution caused by VCR technology over 8 and 16 mm. film technology.

Fiber optics is clearly the enabling technology for high-speed transmission, in fact, so much so that there is an expectation that the switching elements will now hold down the data rates. Both conventional circuit switching and packet switching have significant problems at higher data rates. For instance, circuit switching requires increasing delays for FTDM synchronization to handle skew. In the case of packet switching, traditional approaches require too much processing per packet to handle the tremendous data flow. The problem for both switching regimes is the "intelligence" in the switches, which in turn requires electronics technology.

Besides intelligence, another problem for wide-area networks is storage, both because it ties us to electronics (for the foreseeable future) and because it produces instabilities in a large-scale system. (See, for instance, the work by Van Jacobson on self-organizing phenomena for self-destruction in the Internet.) Techniques are required to eliminate dependence on storage, such as cut-through routing.

Overall, high-speed WANs are the greatest agents of change, the greatest catalyst both commercially and militarily, and the area ripe for revolution. Judging by the attributes of current high-speed network research prototypes, WANs of the future will be photonic, multi-gigabit networks with enormous throughput, low delay, and low error rate.

A zero-based budgeting approach is required to develop the new high-speed internetwork architecture. That is, the time is ripe to significantly rethink the Internet, building on experience with this system. Issues of concern are manageability, understanding evolvability and support for the new communication requirements, including remote procedure call, real-time, security and fault-tolerance.

The GN must be able to deal with two sources of high-bandwidth requirements. There will be some end devices (computers) connected more or less directly to the GN because of their individual requirements for high bandwidth (e.g., supercomputers needing to drive remote high-bandwidth graphics devices). In addition, the aggregate traffic due to large numbers of moderate rate users (estimates are roughly up to a million potential users needing up to 1 Mbit/s at any given time) results in a high-bandwidth requirement in total on the GN. The statistics of such traffic are different and there are different possible technical approaches for dealing with them. Thus, an architectural approach for dealing with both must be developed.

Overall, the next-generation architecture has to be, first and foremost, a management architecture. The directions in link speeds, processor speeds and memory solve the performance problems for many communication situations so well that manageability becomes the predominant concern. (In fact, fast communication makes large systems more prone to performance, reliability, and security problems.) In many ways, the management system of the internetwork is the ultimate distributed system. The solution to this tough problem may well require the best talents from the communications, operating systems and distributed systems communities, perhaps even drawing on database and parallelism research.

3.1.1. High-Speed Internet using High-Speed Networks

The GN will need to take advantage of a multitude of different and heterogeneous networks, all of high speed. In addition to networks based on the technology of the GB, there will be high-speed LANs. A key issue in the development of the GN will be the development of a strategy for interconnecting such networks to provide gigabit service on an end to end basis. This will involve techniques for switching, interfacing, and management (as discussed in the sections below) coupled with an architecture that allows the GN to take full advantage of the performance of the various high-speed networks.

3.1.2. Network Organization

The GN will need an architecture that supports the need to manage the system as well as obtain high performance. We note that almost all human-engineered systems are hierarchically structured from the standpoint of control, monitoring, and information flow. A hierarchical design may be the key to manageability in the next-generation architecture.

One approach is to use a general three-level structure, corresponding to interadministrational, intraadministrational, and cluster networks. The first level interconnects communication facilities of truly separate administrations where there is significant separation of security, accounting, and goals. The second level interconnects subadministrations which exist for management convenience in large organizations. For example, a research group within a university may function as a subadministration. The cluster level consists of networks configured to provides maximal performance among hosts which are in frequent communication, such as a set of diskless workstations and their common file server. These hosts are typically, but not necessarily, geographically collocated. For example, two remote networks may be tightly coupled by a fiber optic link that bridges between the two physical networks, making them function as one.

Research along these lines should study the interorganizational characteristics of communications, such as those being investigated by the IAB Task Force on Autonomous Networks. Based on current results, we expect that such work would clearly demonstrate that considerable communication takes place between particular subadministrations in different administrations; communication patterns are not strictly hierarchical. For example, there might be intense direct communication between the experimental physics departments of two independent universities, or between the computer support group of one company and the operating system development group of another. In addition, (sub)administrations may well also require divisions into public information and private information.

3.1.3. Fault-Tolerant System

Although the GN will be developed as part of an experimental research program, it will also serve as part of the infrastructure for researchers who are experimenting with applications which will use such a network. The GN must have reasonably high availability to support these research activities. In addition to facilitate the transfer of this technology to future operational military and

commercial users, it will need to be designed to become highly reliable. This can be accomplished through diversity of transmission paths, the development of fault-tolerant switches, use of a distributed control structure with self-correcting algorithms, and the protection of network control traffic. The architecture of a GN should support and allow for all of these things.

3.1.4. Functional Division of Control Between Network Elements

Current protocol architectures use the layered model of functional decomposition first developed in the early work on ARPANET protocols. The concept of layering has been a powerful concept which has allowed dramatic variation in network technologies without requiring the complete reimplementation of applications. The concept of layering has had a first-order impact on the development of international standards for data communication---witness the ISO "Reference Model for Open Systems Interconnection."

Unfortunately, however, the powerful concept of layering has been paired, both in the DoD Internet work and the ISO work, with an extremely weak concept of the interface between layers. The interface designs are all organized around the idea of commands and responses plus an error indicator. For example, the TCP service interface provides the user with commands to set up or close a TCP connection and commands to send and receive datagrams. The user may well "know" whether they are using a file transfer service or a character-at-a- time virtual terminal, but can't tell the TCP. The underlying network may "know" that failures have reduced the path to the user's destination to a single 9.6 kbit/s link, but it also can't tell the TCP implementation.

All of the information that an analyst would consider crucial in diagnosing system performance is carefully hidden from adjacent layers. One "solution" often discussed (but rarely implemented) is to condense all of this information into a few bits of "Type of Service" or "Quality of Service" request flowing in one direction only---from application to network. It seems likely that this approach cannot succeed, both because it applies too much compression to the knowledge available and because it does not provide two-way flow.

We believe it to be likely that the next-generation network will require a much richer interface between every pair of adjacent layers if adequate performance is to be achieved. Research is needed into the conceptual mechanisms, both indicators and controls, that can be implemented at these interfaces and that, when used, will result in

better performance. If real differences in performance can be observed, then the implementors of every layer will have a strong incentive to make use of the mechanisms.

We can observe the first glimmers of this sort of coordination between layers in current work. For example, in the ISO work there are 5 classes of transport protocol which are supposed to provide a range of possible matches between application needs and network capabilities. Unfortunately, it is the case today that the class of transport protocol is chosen statically, by the implementer, rather than dynamically. The DARPA Wideband net offers a choice of stream or datagram service, but typically a given host uses all one or all the other---again, a static rather than a dynamic choice. The research that we believe is needed, therefore, is not how to provide alternatives, but how to provide them and choose among them on a dynamic, real-time basis.

3.1.5. Different Switch Technologies

One approach to high-performance networking is to design a technology that is expected to work as a stand-alone demonstration, without addressing the need for interconnection to other networks. Such an experiment may be very valuable for rapid exploration of the design space. However, our experience with the Internet project suggests that a primary research goal should be the development of a network architecture that permits the interconnection of a number of different switching technologies.

The Internet project was successful to a large extent because it could incorporate a number of new and preexisting network technologies: various local area networks, store and forward switching networks, broadcast satellite nets, packet radio networks, and so on. In this way, it decoupled the use of the protocols from a particular technology base. In fact, the technology base evolved rapidly, but the Internet protocols themselves provided a stability that led to their success.

The next-generation architecture must similarly deal with a diverse and evolving technology base. We see "fast-packet" switching now being developed (for example in B-ISDN); we see photonic switching and wavelength division multiplexing as more advanced technologies. We must divorce our architecture from dependence on any one of these.

At the host interface, we must divorce the multiplexing of the medium from the form of data that the host sees. Today the packet is used both as multiplexing and interface element. In the future, the host

may see the network as a message-passing system, or as memory. At the same time, the network may use classic packets, wavelength division, or space division switching.

A number of basic functions must be rethought to provide an architecture that is not dependent on the underlying switching model. For example, our transport protocols assume that data will be lost in units of a packet. If part of a packet is lost, we discard the whole thing. And if several packets are systematically lost in sequence, we may not recover effectively. There must be a host-level unit of error recovery that is independent of the network. This sort of abstraction must be applied to all the aspects of service specification: error recovery, flow control, addressing, and so on.

3.1.6. Network Operations, Monitoring, and Control

There is a hierarchy of progressively more effective and sophisticated techniques for network management that applies regardless of network bandwidth and application considerations:

1. Reactive problem management
2. Reactive resource management
3. Proactive problem management
4. Proactive resource management.

Today's network management strategies are primarily reactive rather than proactive: Problem management is initiated in response to user complaints about service outages; resource allocation decisions are made when users complain about deterioration of quality of service. Today's network management systems are stuck at step 1 or perhaps step 2 of the hierarchy.

Future network management systems will provide proactive problem management---problem diagnosis and restoral of service before users become aware that there was a problem; and proactive resource management---dynamic allocation of network bandwidth and switching resources to ensure that an acceptable level of service is continuously maintained.

The GN management system should be expected to provide proactive problem and resource management capabilities. It will have to do so while contending with three important changes in the managed network environment:

1. More complicated devices under management

2. More diverse types of devices

3. More variety of application protocols.

Performance under these conditions will require that we seriously re-think how a network management system handles the expected high volumes of raw management-related data. It will become especially important for the system to provide thresholding, filtering, and alerting mechanisms that can save the human operator from drowning in data, while still permitting access to details when diagnostic or fault isolation modes are invoked.

The presence of expert assistant capabilities for early fault detection, diagnosis, and problem resolution will be mandatory. These capabilities are highly desirable today, but they will be essential to contend with the complexity and diversity of devices and applications in the Gigabit Network.

In addition to its role in dealing with complexity, automation provides the only hope of controlling and reducing the high costs of daily management and operation of a GN.

Proactive resource management in GNs must be better understood and practiced, initially as an effort requiring human intervention and direction. Once this is achieved, it too must become automated to a high degree in the GN.

3.1.7. Naming and Addressing Strategies

Current networks, both voice (telephone) and data, use addressing structures which closely tie the address to the physical location on the network. That is, the address identifies a physical access point, rather than the higher-level entity (computer, process, human) attached to that access point. In future networks, this physical aspect of addressing must be removed.

Consider, for example, finding the desired party in the telephone network of today. For a person not at his listed number, finding the number of the correct telephone may require preliminary calls, in which advice is given to the person placing the call. This works well when a human is placing the call, since humans are well equipped to cope with arbitrary conversations. But if a computer is placing the call, the process of obtaining the correct address will have to be incorporated in the architecture as a core service of the network.

Since it is reasonable to expect mobile hosts, hosts that are connected to multiple networks, and replicated hosts, the issue of mapping to the physical address must be properly resolved.

To permit the network to maintain the dynamic mapping to current physical address, it is necessary that high-level entities have a name (or logical address) that identifies them independently of location. The name is maintained by the network, and mapped to the current physical location as a core network service. For example, mobile hosts, hosts that are connected to multiple networks, and replicated hosts would have static names whose mapping to physical addresses (many-to-one, in some cases) would change with time.

Hosts are not the only entities whose physical location varies. Users' electronic mail addresses change. Within distributed systems, processes and files migrate from host to host. In a computing environment where robustness and survivability are important, entire applications may move about, or they may be redundant.

The needed function must be considered in the context of the mobility and address resolution rates if all addresses in a global data network were of this sort. The distributed network directory discussed elsewhere in this report should be designed to provide the necessary flexibility, and responsiveness. The nature and administration of names must also be considered.

Names that are arbitrary or unwieldy would be barely better than the addresses used now. The name space should be designed so that it can easily be partitioned among the agencies that will assign names. The structure of names should facilitate, rather than hinder, the mapping function. For example, it would be hard to optimize the mapping function if names were flat and unstructured.

3.2. High-Speed Switching

The term "high-speed switching" refers to changing the switching at a high rate, rather than switching high-speed links, because the latter is not difficult at low speeds. (Consider, for example, manual switching of fiber connections). The switching regime chosen for the network determines various aspects of its performance, its charging policies, and even its effective capabilities. As an example of the latter, it is difficult to expect a circuit-switched network to provide strong multicast support.

A major area of debate lies in the choice between packet switching and circuit switching. This is a key research issue for the GN,

considering also the possibility of there being combinations of the two approaches that are feasible.

3.2.1. Unit of Management vs. Multiplexing

With very high data rates, either the unit of management and switching must be larger or the speed of the processor elements for management and switching must be faster. For example, at a gigabit, a 576 byte packet takes roughly 5 microseconds to be received so a packet switch must act extremely fast to avoid being the dominant delay in packet times. Moreover, the storage time for the packet in a conventional store and forward implementation also becomes a significant component of the delay. Thus, for packet switching to remain attractive in this environment, it appears necessary to increase the size of packets (or switch on packet groups), do so-called virtual cut-through and use high-speed routing techniques, such as high-speed route caches and source routing.

Alternatively, for circuit switching to be attractive, it must provide very fast circuit setup and tear-down to support the bursty nature of most computer communication. This problem is rendered difficult (and perhaps impossible for certain traffic loads) because the delay across the country is so large relative to the data rate. That is, even with techniques such as so-called fast select, bandwidth is reserved by the circuit along the path for almost twice the propagation time before being used.

With gigabit circuit switching, because it is not feasible to physically switch channels, the low-level switching is likely doing FTDM on micro-packets, as is currently done in telephony. Performing FTDM at gigabit data rates is a challenging research problem if the skew introduced by wide-area communication is to be handled with reasonable overhead for spacing of this micro-packets. Given the lead and resources of the telephone companies, this area of investigation should, if pursued, be pursued cooperatively.

3.2.2. Bandwidth Reservation Algorithms

Some applications, such as real-time video, require sustained high data rate streams over a significant period of time, such as minutes if not hours. Intuitively, it is appealing for such applications to pre-allocate the bandwidth they require to minimize the switching load on the network and guarantee that the required bandwidth is available. Research is required to determine the merits of bandwidth

reservation, particular in conjunction with the different switching technologies. There is some concern to raise that bandwidth reservation may require excessive intelligence in the network, reducing the performance and reliability of the network. In addition, bandwidth reservation opens a new option for denial of service by an intruder or malicious user. Thus, investigations in this area need to proceed in concert with work on switching technologies and capabilities and security and reliability requirements.

3.2.3. Multicast Capabilities

It is now widely accepted that multicast should be provided as a user-level service, as described in RFC 1054 for IP, for example. However, further research is required to determine the best way to support this facility at the network layer and lower. It is fairly clear that the GN will be built from point-to-point fiber links that do not provide multicast/broadcast for free. At the most conservative extreme, one could provide no support and require that each host or gateway simulate multicast by sending multiple, individually addressed packets. However, there are significant advantages to providing very low level multicast support (besides the obvious performance advantages). For example, multicast routing in a flooding form provides the most fault-tolerant, lowest-delay form of delivery which, if reserved for very high priority messages, provides a good emergency facility for high-stress network applications. Multicast may also be useful as an approach to defeat traffic analysis.

Another key issue arises with the distinction between so-called open group multicast and closed group multicast. In the former, any host can multicast to the group, whereas in the latter, only members of the group can multicast to it. The latter is easier to support and adequate for conferencing, for example. However, for more client-server structured applications, such as using file/database server, computation servers, etc. as groups, open multicast is required. Research is needed to address both forms of multicast. In addition, security issues arise in controlling the membership of multicast groups. This issue should be addressed in concert with work on secure forms of routing in general.

3.2.4. Gateway Technologies

With the wide-area interconnection of local networks by the GN, gateways are expected to become a significant performance bottleneck unless significant advances are made in gateway performance. In addition, many network management concerns suggest putting more functionality (such as access control) in the gateways, further increasing their load and the need for greater capacity. This would then raise the issue of the trade-off between general-purpose hardware and special-purpose hardware.

On the general-purpose side, it may be feasible to use a general-purpose multiprocessor based on high-end microprocessors (perhaps as exotic as the GaAs MIPS) in conjunction with a high-speed block transfer bus, as proposed as part of the FutureBus standard (which is extendible to higher speeds than currently commercially planned) and intelligent high-speed network adaptors. This would also allow the direct use of hardware, operating systems, and software tools developed as part of other DARPA programs, such as Strategic Computing. It also appears to make this gateway software more portable to commercial machines as they become available in this performance range.

The specialized hardware approach is based on the assumption that general-purpose hardware, particularly the interconnection bus, cannot be fast enough to support the level of performance required. The expected emphasis is on various interconnection network techniques. These approaches appear to require greater expense, less commercial availability and more specialized software. They need to be critically evaluated with respect to the general-purpose gateway hardware approach, especially if the latter is using multiple buses for fault-tolerance as well as capacity extension (in the absence of failure).

The same general-purpose vs. special-purpose contention is an issue with operating system software. Conventionally, gateways run specialized run-time executives that are designed specifically for the gateway and gateway functions. However, the growing sophistication of the gateway makes this approach less feasible. It appears important to investigate the feasibility of using a standard operating system foundation on the gateways that is known to provide the required security and reliability properties (as well as real-time performance properties).

3.2.5. VLSI and Optronics Implementations

It appears fairly clear that gigabit communication will use fiber optics for at least the near future. Without major advances in optronics to allow effectively for optical computers, communication must cross the optical-electronic boundary two or more times. There are significant cost, performance, reliability, and security benefits for minimizing the number of such crossings. (As an example of a security benefit, optics is not prone to electronic surveillance or jamming while electronics clearly is, so replacing an optic-electronic-optic node with a pure optic node eliminates that vulnerability point.)

The benefits of improved technology in optronics is so great that its application here is purely another motivation for an already active research area (that deserves strong continued support). Therefore, we focus here in the issue of matching current (and near-term expected) optronics capabilities with network requirements.

The first and perhaps greatest area of opportunity is to achieve totally (or largely) photonic switches in the network switching nodes. That is, most packets would be switched without crossing the optics-electronics boundary at all. For this to be feasible, the switch must use very simple switching logic, require very little storage and operate on packets of a significant size. The source-routed packet switches with loopback on blockage of Blazenet illustrate the type of techniques that appear required to achieve this goal.

Research is required to investigate the feasibility of optronic implementation of switches. It appears highly likely that networks will at some point in the future be totally photonically switched, having the impact on networking comparable to the effect of integrated circuits on processors and memories.

A next level of focus is to achieve optical switching in the common case in gateways. One model is a multiprocessor with an optical interconnect. Packets associated with established paths through the gateway are optically switched and processed through the interconnect. Other packets are routed to the multiprocessor, crossing into the electronics domain. Research is required to marry the networking requirements and technology with optronics technology, pushing the state of the art in both areas in the process.

Given the long-term presence of the optic-electronic boundary, improvements in technology in this area are also important. However, it appears that there is already enormous commercial research

activity in this area, particularly within the telephone companies. This is another area in which collaborative investigation appears far better than an new independent research effort.

VLSI technology is an established technology with active research support. The GN effort does not appear to require major new initiatives in the VLSI area, yet one should be open to significant novel opportunities not identified here.

3.2.6. High-Speed Transfer Protocols

To achieve the desired speeds, it will be necessary to rethink the form of protocols.

1. The simple idea of a stateless gateway must be replaced by a more complex model in which the gateway understands the desired function of the end point and applies suitable optimizations to the flow.

2. If multiplexing is done in the time domain, the elements of multiplexing are probably so small that no significant processing can be performed on each individually. They must be processed as an aggregate. This implies that the unit of multiplexing is not the same as the unit of processing.

3. The interfaces between the structural layers of the communication system must change from a simple command/response style to a richer system which includes indications and controls.

4. An approach must be developed that couples the memory management in the host and the structure of the transmitted data, to allow efficient transfers into host memory.

The result of rethinking these problems will be a new style of communications and protocols, in which there is a much higher degree of shared responsibility among the components (hosts, switches, gateways). This may have little resemblance to previous work either in the DARPA or commercial communities.

3.3. High-Speed Host Interfaces

As networks get faster, the most significant bottleneck will turn out to be the packet processing overhead in the host. While this does

not restrict the aggregate rates we can achieve over trunks, it prevents delivery of high data rate flows to the host-based applications, which will prevent the development of new applications needing high bandwidth. The host bottleneck is thus a serious impediment to networked use of supercomputers.

To build a GN we need to create new ways for hosts and their high bandwidth peripherals to connect to networks. We believe that pursuing research in the ways to most effectively isolate host and LAN development paths from the GN is the most productive way to proceed. By decoupling the development paths, neither is restricted by the momentary performance of capability bottlenecks of the other. The best context in which to view this separation is with the notion of a network front end (NFE). The NFE can take the electronic input data at many data rates and transform it into gigabit light data appropriately packetized to traverse the GN. The NFE can accept inputs from many types of gateways, hosts, host peripherals, and LANS and provide arbitration and path set-up facilities as needed. Most importantly, the NFE can perform protocol arbitration to retain upward compatibility with the existing Internet protocols while enabling those sophisticated network input sources to execute GN specific high-throughput protocols. Of course, this introduces the need for research into high-speed NFEs to avoid the NFE becoming a bottleneck.

3.3.1. VLSI and Optronics Implementations

In a host interface, unless the host is optical (an unlikely prospect in the near-term), the opportunities for optronic support are limited. In fact, with a serial-to-parallel conversion on reception stepping the clock rate down by a factor of 32 (assuming a 32-bit data path on the host interface), optronic speeds are not required in the immediate future.

One exception may be for encryption. Current VLSI implementations of standard encryption algorithms run in the 10 Mbit/s range. Optronic implementation of these encryption techniques and encryption techniques specifically oriented to, or taking advantage of, optronic capabilities appears to be an area of some potential (and enormous benefit if achieved).

The potential of targeted VLSI research in this area appears limited for similar reasons discussed above with its application in high-speed switching. The major benefits will arise from work that is well-motivated by other research (such as high-performance parallelism) and by strong commercial interest. Again, we need to be

open to imaginative opportunities not foreseen here while keeping ourselves from being diverted into low-impact research without further insights being put forward.

3.3.2. High-Performance Transport Protocols

Current transport protocols exhibit some severe problems for maximal performance, especially for using hardware support. For example, TCP places the checksum in the packet header, forcing the packet to be formed and read fully before transmission begins. ISO TP4 is even worse, locating the checksum in a variable portion of the header at an indeterminate offset, making hardware implementation extremely difficult.

The current Internet has thrived and grown due to the existence of TCP implementations for a wide variety of classes of host computers. These various TCP implementations achieve robust interoperability by a "least common denominator" approach to features and options. Some applications have arisen in the current Internet, and analogs can be envisioned for the GN environment, which need qualities of service not generally supported by the ubiquitous generic TCP, and therefore special purpose transport protocols have been developed. Examples include special purpose transport protocols such as UDP (user datagram protocol), RDP (reliable datagram protocol), LDP (loader/debugger protocol), NETBLT (high-speed block transfer protocol), NVP (network voice protocol) and PVP (packet video protocol). Efforts are also under way to develop a new generic transport protocol VMTP (versatile message transaction protocol) which will remedy some of deficiencies of TCP, without the need to resort to special purpose protocols for some applications. Research is needed in this area to understand how transport level protocols should be constructed for a GN which provide adequate qualities of service and ease of implementation.

A new transport protocol of reasonable success can be expected to last for ten years more. Therefore, a new protocol should not be over optimized for current networks and must not ignore the functional deficiencies of current protocols. These deficiencies are essential to remedy before it is feasible to deploy even current distributed systems technology for military and commercial applications.

Forward Error Correction (FEC) is a useful approach when the bandwidth/delay ratio of the physical medium is high, as can be expected in transcontinental photonic links. A degenerate form of FEC is to simply transmit multiple copies of the data; this allows

one to trade bandwidth for delay and reliability, without requiring much intelligence. In fact, it is generally true that reliability, bandwidth, and delay are interrelated and an improvement in one generally comes at the expense of the others for a given technology. Research is required to find appropriate operating points in networks using transmission components which offer extremely high bandwidth with very good bit-error-rate performance.

3.3.3. Network Adaptors

With the promised speed of networks, the future network adaptor must be viewed as a memory interconnect, tying the memory in one host to another, at least if the data rate and the low latency made possible by the network is to be realized at the host-to-host or process-to-process level. The challenge is too great to be met by just implementing protocols in custom VLSI.

Research is required to investigate the impact of network interconnection on a machine architecture and to define and evaluate new network adaptor architectures. Of key importance is integration of network adaptor into the operating system so that process-to-process communications performance matches that offered by the network. In particular, we conjecture that the transport level will be implemented largely, if not entirely, in the network adaptor, providing the host with reliable memory-to-memory transfer at memory speeds with a minimum of interrupt processing bus overhead and packet processing.

Drawing an analogy to RISC technology again, maximal performance requires a well-designed and coordinated protocol, software, and hardware (network adaptor) design. Current standard protocols are significantly flawed for hardware compatibility, suggesting a need for considerable further research on high-performance protocol design.

3.3.4. Host Operating System Software

Conventionally, communication has been an add-on to an operating system. With the GN, the network may well become the fastest "peripheral" connected to most nodes. High-performance process-to-process (or application to application) communication will not be achieved until the operating system is well designed for fast access to and from the network. For example, incorporating templates of the network packet header directly in the process descriptor may allow a

process to initiate communications with minimal overhead. Similarly, memory mapping can be used to eliminate copies between data arriving from the network and it being delivered to the applications. With a GN, an extra copy forced by the operating system may easily double the perceived transfer time for a packet between applications.

Besides matching data transfer mechanisms, operating systems must be well-matched in security design to that supported by the host interface and network as well. Otherwise, all but the most trivial additional security actions by the operating system in common case communication can easily eliminate the performance benefits of the GN. For example, if the host has to do further encryption or decryption, the throughput is likely to be at least halved and the latency doubled.

Research effort is required to further refine operating systems for the level of performance offered by the GN. This effort may well be best realized with coupling existing efforts in distributed systems with the GN activities, as opposed to starting new separate efforts.

3.4. Advanced Network Management Algorithms

An important emphasis for research into network management should be on decentralized approaches. The ratio of propagation delay across the country to data rates in a GN appear to be too great to deal effectively with resource management centrally when traffic load is bursty and unstable (and if it is not, one might argue there is no problem). In addition, important principles of fault containment and minimal privilege for reliability and security suggest that a centralized management approach is infeasible. In particular, compromising the security of one portion of the network should not compromise the security of the whole network. Similarly, a failure or fault should affect at most a local region of the network.

The challenge is clearly to provide decentralized management techniques that lead to good global behavior in the normal case and acceptable behavior in expected worst-case failures, traffic variations and security intrusions.

3.4.1. Control Flow vs. Data Flow

Network operational communications can be separated into flow of user data and flow of management/control data. However, the user data must contain some amount of control data. One question that needs to

be explored in light of changes in communications and computing costs and performance is the trade-off between these two flows. An example of a potential approach is to use data units which contain predefined path indicators. The switch can perform a simple table look-up which maps the path indicator onto the preferred outbound link and transmits the packet immediately. There is a path set-up packet which fills in the appropriate tables. Path set-up occurs before the first data packet flows and then, while data is flowing, to improve the routes during the lifetime of the connection. This concept has been discussed in the Internet engineering group under the name of soft connections.

We note that separating the data flow from the control flow in the GN has security and reliability advantages as well. We could encrypt most of the packet header to provide confidentiality within the GN and to limit the ability of intruders to perform traffic analysis. And, by separating the control flow, we can encrypt all the control exchanges between switches and the host front ends thereby offering confidentiality and integrity. No unauthorized entity will be able to alter or examine the control traffic. By employing a path set-up procedure, we can assure that the GN NFE-to-NFE path is functioning and also include user-specific requirements in the route. For example, we could request a certain bandwidth allocation and simplify the job of the switches in handling flow control. We could also set up backup paths in case the output link will be busy for so many microseconds that the packet cannot be stored until the link is freed.

3.4.2. Resource Management Algorithms

Most current networks deliver one quality of service. X.25 networks deliver a reliable byte-stream. Most LANs deliver a best-effort unreliable service. There are few networks today that can support multiple types of service, and allocate their resources among them. Indeed, for many networks, such as best-effort unreliable service, there is little management of resources at all. The next generation of network will require a much more controlled allocation of resources.

There will be a much wider range of desired types of service, with current services such as remote procedure call mixing with new services such as video streams. Unless these are separately recognized and controlled, there is little reason to believe that effective service can be delivered unless the network is very lightly loaded.

In order to support multiple types of service, two things must happen, both a change from current practice. First, the application must describe to the network what type of service is required. Second, the network must use this information to make resource allocation decisions. Both of these practices present difficulties.

Past experience suggests that application code is not prepared to know or specify what service it needs. By custom, operating systems provide a virtual world, and the applications in this world are unaware of the relation between this and the reality of time and space. Resource requests must be in real terms. Allocation of resources in the network is difficult, because it requires that decisions be made in the network, but as network packet throughput increases, there is less time for decisions.

The resolution of this latter conflict is to observe that decisions must be made on larger units than the unit of multiplexing such as the packet. This in turn implies that packets must be visible to the network as being part of a sequence, as opposed to the pure datagram model previously exploited. As suggested earlier in this report, research is required to support this more complex form of switch without compromising robustness.

To permit the application to specify the service it needs, it will be necessary to propose some abstraction of service class. By clever design of this abstraction, it should be possible to allow the application to describe its needs effectively. For example, an application such as file transfer or mail has two modes of operation; bulk data transfer and remote procedure call. The application may not be able to predict when it will be in which mode, but if it just describes both of them, the system may be able to adapt by observing its current operation.

Experimentation needs to be done to determine a suitable service specification interface. This experimentation could be done in the context of the current protocols, and could thus be undertaken at once.

3.4.3. Adaptive Protocols

Network operating conditions can vary quickly and over a wide range. This is true of the current Internet, and is likely to affect the GN too. Protocols that can adapt to changing circumstances would provide more even and robust service than is currently possible. For example, when error rates increased, a protocol implementation might decide to use smaller packets, thus reducing the burden caused by

retransmissions.

The environment in which a protocol operates can be described in terms of the service it is getting from the next lower layer. A protocol implementation can adapt to changes in that service by tuning its internal mechanisms (time-outs, retransmission strategies, etc.). Therefore, to design adaptive protocols, we must understand the interaction between protocol layers and the mechanisms used within them. There has been some work done in this area. For example, the SATNET measurement task force has looked at the interactions between the protocol used by the SIMP, IP, and TCP. What is needed is a more complete characterization of the interactions at various layer boundaries, and the development of appropriate protocol designs and mechanisms to provide for necessary adaptations and renegotiations.

3.4.4. Error Recovery Mechanisms

Being large and complex, GNs will experience a variety of faults such as link or nodal failure, excessive buffer overflow due to faulty flow and congestion control, and partial failure of switching fabric. These failures, which also exist in today's networks, will have a stronger effect in GNs where a large amount of data will be "stored" in transit and, to expedite the switching, nodes will apply only minimal processing to the packets traversing them. In source routing, for example, a link failure may cause the loss of all packets sent until the source is notified about the change in topology. The longer is the delay in recovering from failures, the higher is the degradation in performance observed by the users.

To minimize the effects of failures, GNs will need to employ error recovery mechanisms whereby the network detects failures and error conditions, reconfigures itself to adapt to the new network state, and notifies peripheral devices of the new configuration. Such protocols, which have to be developed, will respond quickly, will be decentralized or distributed to minimize the possibility of fatal failures, and will complement, rather than replicate, the error correction mechanisms of the end-to-end protocols, and the two must operate in coordinated manner. To this end, the peripheral devices will have to be knowledgeable about the intranet recovery mechanisms and interact continuously with them to minimize the effect on the connections they manage.

3.4.5. Flow Control

As networks become faster, two related problems arise. First, existing flow control mechanisms such as windows do not work well, because the window must be opened to such an extent to achieve desired bandwidth that effective flow control cannot be achieved. Second, especially for long-haul networks, the larger number of bits in transit at one time becomes so large that most computer messages will fit into one window. This means that traditional congestion control schemes will cease to work well.

What is needed is a combination of two approaches, both new. First, for messages that are small (most messages generated by computers today will be small, since they will fit into one round-trip time of future networks), open-loop controls on flow and congestion are needed. For longer messages (voice or video streams, for example), some explicit resource commitment will be required.

3.4.6. Latency Control and Real-Time Operations

Currently, there are several distinct approaches to latency control. First, there are some networks which are physically short, more like multiprocessor buses. Applications in these networks are built assuming that delays will be short.

Second, there are networks where the physical length is not constrained by the design and may differ by orders of magnitude, depending on the scope of the network. Most general purpose networks fall in this category. In these networks, one of two things happens. Either the application takes special steps to deal with variable latency, such as echo suppression in voice networks, or these applications are not supported.

For most applications today, the latency in the network is not an obvious issue so long as the network is not overloaded (which leads to losses and long queues), because the protocol overhead masks the variation in the network latency. This balance will change. The latency due to the speed of light will obviously remain the same, but the overhead will drop (of necessity if we are to achieve high performance) which will leave speed of light and queueing as the most critical sources of delay.

This conclusion implies that if queueing delay can be controlled, it will be possible to build networks with stable and controlled latency. If applications exist that require this class of service,

it can be supported. Either the network must be underloaded, so that queues do not develop at all, or a specific class of service must be supported in which resources are allocated to stabilize the delay.

If this service is provided, it will still leave the application with delays that can vary by several orders of magnitude, depending on the physical size of the network. Research at the application level will be required to see how applications can be designed to cope with this variation.

3.4.7. High-Speed Internetworking and Administrational Domains

Internetworking recognized that the value of communication services increases significantly with wider interconnection but ignored management and the role of administrations. As a consequence we see that:

1. The Internet is more or less unmanageable, as evidenced by performance, reliability, and security problems.

2. The Internet is being stressed by administrators that are building networks to match their organization rather than the geography. An example is a set of Ethernets at different company locations operating as a single Internet network but geographically dispersed and connected by satellite or leased lines.

The next generation of internetworking must focus on administration and management. Internetworking must support cohesion within an administration and a healthy separation between administrations. To illustrate by analogy, the American and Soviet embassies in Mexico City are geographically closer to each other than to their respective home countries but further in administrational distance, including security, accounting, etc. The emerging revolution in WANs makes this issue that much more critical. The amount of communication to exchange the state of systems is bound to increase enormously. The potential cost of failures and security violations is frightening.

A promising approach appears to be high-level gateways that guard between administrations and require negotiations to set up access paths between administrations. These paths are set up, and labeled with agreements on authorization, security, accounting, and possible resource limits. These administrative virtual circuits provide transparency to the physical and geographical interconnection, but need not support more than datagram packet delivery. One view is that of communication contracts with high-level gateways acting as

contract monitors at each end. The key is the focus on controlled interadministrational connectivity, not the conventional protocol concerns.

Focus is required on developing an (inter)network management architecture and the specifics of high-level gateways. The structures of such gateways will have to take advantage of advances in multi-processor architectures to handle the processing load. Moreover, a key issue is being able to optimize communication between administrations once the contract is in place, but without losing control. Related is the issue of allowing high-speed interconnection within a single administration, although geographical dispersed. Another issue is fault-tolerance. High-level gateways contain state information whose loss typically disrupts communication. How does one minimize this problem?

A key goal of these administrational gateways has to be failure containment: How to protect against external (to administration) problems and how to prevent local problems imposing liability on others. A particular area of concern is the self-organizing problems of large-scale systems, observed by Van Jacobson in the Internet. Gateways must serve to damp out oscillations and control wide load swings. Rate control appears to be a key area to investigate as a basis for buffer management and for congestion control, as well as to control offered load.

Given the speed of new networks, and the sophistication of the gateways suggested above, another key area to investigate is the provision of high-speed network interface adaptors.

3.4.8. Policy-Based Algorithms

Networks of today generally select routes based on minimizing some measure such as delay. However, in the real world, route selection will commonly be constrained at the global level by policy issues, such as access rights to resources and accounting and billing for usage.

It is difficult for connectionless protocols such as Internet to deal with policy controls, because a lack of state in the gateway implies that a separate policy decision must be made for each packet in isolation. As networks get faster, the cost of this processing will be intolerable. One possible approach, discussed above, is to move to a more sophisticated model in which there is knowledge in the gateways of the ongoing flows. Alternatively, it may be possible to design gateways that simply cache recent policy evaluations and apply

them to successive packets.

Routing based on policy is particularly difficult because a route must be globally consistent to be useful; otherwise it may loop. This implies that the every policy decision must be propagated globally. Since there can be expected to be a large number of policies, this global passing of information might easily lead to an information explosion.

There are at least two solutions. One is to restrict the possible classes of policy. Another is to use some form of source route, so that the route consistent with some set of policies is computed at one point only, and then attached to the packet. Both of these approaches have problems. A two-pronged research program is needed, in which mechanisms are proposed, and at the same time the needed policies are defined.

The same trade-off can be seen for accounting and billing. A single accounting metric, such as "bytes times distance", could be proposed. This might be somewhat simple to implement, but would not permit the definition of individual billing policies, as is now done in the parts of the telephone system. The current connectionless transport architectures such as TCP/IP or the connectionless ISO configuration using TP4 do not have good tools for accounting for traffic, or for restricting traffic from certain resources. Building these tools is difficult in a connectionless environment, because an accounting or control facility must deal with each packet in isolation, which implies a significant processing burden as part of packet forwarding. This burden is an increasing problem as switches are expected to operate faster.

The lack of these tools is proving a significant problem for network design. Not only are accounting and control needed to support management requirements, they are needed as a building block to support enforcement of such things as multiple qualities of service, as discussed above.

Network accounting is generally considered to be simply a step that leads to billing, and thus is often evaluated in terms of how simple or difficult it will be to implement. Yet an accounting and billing procedure is a mechanism for implementing a policy considered to be desirable for reasons beyond the scope of accounting per se. For example, a policy might be established either to encourage or discourage network use, while fully recovering operational cost. A policy of encouraging use could be implemented by a relatively high monthly attachment charge and a relatively low per-packet charge. A policy of discouraging use could be implemented by a low monthly charge and a high per-packet charge.

Network administrators have a relatively small number of variables with which to implement policy objectives. Nevertheless, these variables can be combined in a number of innovative ways. Some of the possibilities include:

1. Classes of users (e.g., large or small institutions, for-profit or non-profit).

2. Classes of service.

3. Time varying (e.g., peak and off-peak).

4. Volume (e.g., volume discounts, or volume surcharges).

5. Access charges (e.g., per port, or port * [bandwidth of port]).

6. Distance (e.g., circuit-miles, airline miles, number of hops).

Generally, an accounting procedure can be developed to support voluntary user cooperation with almost any single policy objective. Difficulties most often arise when there are multiple competing policy objectives, or when there is no clear policy at all.

Another aspect of accounting and billing procedures which must be carefully considered is the cost of accumulating and processing the data on which billing is based. Of particular concern is collection of detailed data on a per-packet basis. As network circuit data rates increase, the number of instructions which must be executed on a per-packet basis can become the limiting factor in system throughput. Thus, it may be appropriate to prefer accounting and billing policies and procedures which minimize the difficulty of collecting data, even if this approach requires a compromise of other objectives. Similarly, node memory required for data collection and any network bandwidth required for transmission of the data to administrative headquarters are factors which must be traded off against the need to process user packets.

3.4.9. Priority and Preemption

The GN should support multiple levels of priority for traffic and the preemption of network resources for higher priority use. Network control traffic should be given the highest priority to ensure that it is able to pass through the network unimpeded by congestion caused by user-level traffic. There may be additional military uses for multiple levels of priority which correspond to rank or level of

importance of a user or the mission criticality of some particular data.

The use of and existence of priority levels may be different for different types of traffic. For example, datagram traffic may not have multiple priority levels. Because the network's transmission speed is so high and traffic bursts may be short, it may not make sense to do any processing in the switches to deal with different priority levels. Priority will be more important for flow- (or soft-connection-) oriented data or hard connections in terms of permitting higher priority connections to be set up ahead of lower priority connections. Preemption will permit requests for high priority connections to reclaim network resources currently in use by lower priority traffic.

Networks such as the Wideband Satellite Network, which supports datagram and stream traffic, implement four priority levels for traffic with the highest reserved for network control functions and the other three for user traffic. The Wideband Network supports preemption of lower priority stream allocations by higher priority requests. An important component of the use of priority and preemption is the ability to notify users when requests for service have been denied, or allocations have been modified or disrupted. Such mechanisms have been implemented in the Wideband Network for streams and dynamic multicast groups.

Priority and preemption mechanisms for a GN will have to be implemented in an extremely simple way so that they can take effect very quickly. It is likely that they will have to built into the hardware of the switch fabric.

3.5. User and Network Services

As discussed in Section 2 above, there will need to be certain services provided as part of the network operation to the users (people) themselves and to the machines that connect to the network. These services, which include such capabilities as white and yellow pages (allowing users to determine what the appropriate network identification is for other users and for network-available computing resources) and distributed fault identification and isolation, are needed in current networks and will continue to be required in the networks of the future. The speed of the GN will serve to accentuate this requirement, but at the same time will allow for new architectures to be put in place for such services. For example, Ethernet speeds in the local environment have allowed for more usable services to be provided.

3.5.1. Impact of High Bandwidth

One issue that will need to be addressed is the impact on the user of such high-bandwidth capabilities. Users are already becoming saturated by information in the modern information-rich environment. (Many of us receive more than 50 electronic mail messages each day, each requiring some degree of human attention.) Methods will be needed to allow users to cope with this ever-expanding access to data, or we will run the risk of users turning back to the relative peace and quiet of the isolated office.

3.5.2. Distributed Network Directory

A distributed network directory can support the user-level directory services and the lower-level name-to-address mapping services described elsewhere in this report. It can also support distributed systems and network management facilities by storing additional information about named objects. For example, the network directory might store node configurations or security levels.

Distributing the directory eases and decentralizes the administrative burdens and provides a more robust and survivable implementation.

One approach toward implementing a distributed network directory would be to base it upon the CCITT X.500/ISO DIS 9594 standard. This avoids starting from ground zero and has the advantage of facilitating interoperability with other communications networks. However, research and development will be required even if this path is chosen.

One area in which research and development are required is in the services supplied by the distributed network directory. The X.500 standard is very general and powerful, but so far specific provisions have been made only for storing information about network users and applications. As mentioned elsewhere, multilevel security is not addressed by X.500, and the approach taken toward authentication must be carefully considered in view of DoD requirements. Also, X.500 assumes that administration of the directory will be done locally and without the need for standardization; this may not be true of GN or the larger national research network.

The model and algorithms used by a distributed network directory constitute a second area of research. The model specified by X.500 must be extended into a framework that provides the necessary flexibility in terms of services, responsiveness, data management

policies, and protocol layer utilization. Furthermore, the internal algorithms and mechanisms of X.500 must be extended in a number of areas; for example, to support redundancy of the X.500 database, internal consistency checking, fuller sharing of information about the distribution of data, and defined access-control mechanisms.

4. Avenues of Approach

Ongoing research and commercial activities provide an opportunity for more rapidly attacking some of the above research issues. At the same time, there needs to be attention paid to the overall technical approach used to allow multiple potential solutions to be explored and allow issues to be attacked in parallel.

4.1. Small Prototype vs. Nationwide Network

The central question is how far to jump, and how far can the current approaches get. That is, how far will connectionless network service get us, how far will packet switching get us, and how far do we want to go. If our goal is a Gbit/s net, then that is what we should build. Building a 100 Mbit/s network to achieve a GN is analogous to climbing a tree to get to the moon. It may get you closer, but it will never get you there.

There are currently some network designs which can serve as the basis for a GN prototype. The next step is some work by experts in photonics and possibly high-speed electronics to explore ease of implementation. Developing a prototype 3-5 node network at a Gbit/s data rate is realistic at this point and would demonstrate wide-area (40 km or more) Gbit/s networking.

DARPA should consider installing a Gbit/s cross-country set of connected links analogous to the NSF backbone in 2 years. A Gbit/s link between the east and west coasts would open up a whole new generation of (C3I), distributed computing, and parallel computing research possibilities and would reestablish DARPA as the premier network research funding agency in the country. This will require getting "dark" fiber from one or more of the common carriers and some collaboration with these organizations on repeaters, etc. With this collaboration, the time to a commercial network in the Gbit/s range would be substantially reduced, and the resulting nationwide GN would give the United States an enormous technical and economic advantage over countries without it.

Demonstrating a high-bandwidth WAN is not enough, however. As one can see from the many research issues identified above, it will be necessary to pursue via study and experiment the issues involved in interconnecting high-bandwidth networks into a high-bandwidth internet. These experiments can be done through use of a new generation of internet, even if it requires starting at lower speeds (e.g., T1 through 100 Mbit/s). Appropriate care must be given, however, to assure that the capabilities that are demonstrated are applicable to the higher bandwidths (Gbit/s) as they emerge.

4.2. Need for Parallel Efforts/Approaches

Parallel efforts will therefore be required for two major reasons. First is the need to pursue alternative approaches (e.g., different strategies for high-bandwidth switching, different addressing techniques, etc). This is the case for most research programs, but it is made more difficult here by the costs of prototyping. Thus, it is necessary that appropriate review take place in the decisions as to which efforts are supported through prototyping.

In addition, it will be necessary to pursue the different aspects of the program in parallel. It will not be possible to wait until the high-bandwidth network is available before starting on prototyping the high-bandwidth internet. Thus, a phased and evolutionary approach will be needed.

4.3. Collaboration with Common Carriers

Computer communication networks in the United States today practically ignore the STN (the Switched Telephone Network), except for buying raw bandwidth through it. However, advances in network performance are based on improvements in the underlying communication media, including satellite communication, fiber optics, and photonic switching.

In the past we used "their" transmission under "our" switching. An alternative approach is to utilize the common-carrier switching capabilities as an integral part of the networking architecture. We must take an objective scientific and economic look and reevaluate this question.

Another place for cooperation with the common carriers is in the area of network addressing. Their addressing scheme ("numbering plan") has a few advantages such as proven service to 300 million users [4].

On the other hand, the common carriers have far fewer administrative domains (area codes) than the current plethora of locally administered local area networks in the internet system.

It is likely that future networks will eventually be managed and operated by commercial communications providers. A way to maximize technology transfer from the research discussed here to the marketplace is to involve the potential carriers from the start. However, it is not clear that the goals of commercial communications providers, who have typically been most interested in meeting the needs of 90+ percent of the user base, will be compatible with the goals of the research described here. Thus, while we recommend that the research program involve an appropriate amalgam of academia and industry, paying particular attention to involvement of the potential system developers and operators, we also caution that the specific and unique goals of the DARPA program must be retained.

4.4. Technology Transfer

As we said above, it is our belief that future networks will ultimately be managed and operated by commercial communications providers. (Note that this may not be the common carriers as we know them today, but may be value-added networks using common carrier facilities.) The way to assure technology transfer, in our belief, is to involve the potential system developers from the start. We therefore believe that the research program would benefit from an appropriate amalgam of university and industry, with provision for close involvement of the potential system developers and operators.

4.5. Standards

The Internet program was a tremendous success in influencing national and international standards. While there were changes to the protocols, the underlying technology and approaches used by CCITT and ISO in the standardization of packet-switched networks clearly had its roots in the DARPA internet. Nevertheless, this has had some negative impact on the research program, as the evolution of the standards led to pressure to adopt them in the research environment.

Thus, it appears that there is a "catch-22" here. It is desirable for the technology base developed in the research program to have maximal impact on the standards activities. This is expedited by doing the research in the context of the standards environment. However, standards by their very nature will always lag behind the

research environment.

The only reasonable approach, therefore, appears to be an occasional "checkpointing" of the research environment, where the required conversions take place to allow a new plateau of standards to be used for future evolution and research. A good example is conducting future research in mail using X.400 and X.500 where possible.

5. Conclusions

We hope that this document has provided a useful compendium of those research issues critical to achieving the FCCSET phase III recommendations. These problems interact in a complex way. If the only goal of a new network architecture was high speed, reasonable solutions would not be difficult to propose. But if one must achieve higher speeds while supporting multiple services, and at the same time support the establishment of these services across administrative boundaries, so that policy concerns (e.g., access control) must be enforced, the interactions become complex.

APPENDIX

A. Current R and D Activities

In this appendix, we provide pointers to some ongoing activities in the research and development community of which the group was aware relevant to the goal of achieving the GN. In some cases, a short abstract is provided of the research. Neither the order of the listing (which is random) nor the amount of detail provided is meant to indicate in any way the significance of the activity. We hope that this set of pointers will be useful to anyone who chooses to pursue the research issues discussed in this report.

1. Grumman (at Bethpage) is working on a three-year DARPA contract, started in January 1988 to develop a 1.6 Gbit/s LAN, for use on a plane or ship, or as a "building block". It is really raw transport capacity running on two fibers in a token-ring like mode. First milestone (after one year?) is to be a 100 Mbit/s demonstration.

2. BBN Laboratories, as part of its current three-year DARPA Network-Oriented Systems contract, has proposed design concepts for a 10-100 Gbit/s wide area network. Work under this effort will include wavelength division multiplexing, photonic switching, self-routing packets, and protocol design.

3. Cheriton (Stanford) research on Blazenet, a high-bandwidth network using photonic switching.

4. Acampora (Bell Labs) research on the use of wavelength division multiplexing for building a shared optical network.

5. Yeh is reserching a VLSI approach to building high-bandwidth parallel processing packet switch.

6. Bell Labs is working on a Metropolitan Area Network called "Manhattan Street Net." This work, under Dr. Maxemchuck, is similar to Blazenet. It is in the prototype stage for a small number of street intersections; ultimately it is meant to be city-wide. Like Blazenet, is uses photonic switching 2 x 2 lithium niobate block switches.

7. Ultra Network Technologies is a Silicon Valley company which has a (prototype) Gbit/s fiber link which connects backplanes. This is based on the ISO-TP4 transport protocol.

8. Jonathan Turner, Washington University, is working on a Batcher-Banyan Multicast Net, based on the "SONET" concept,

which provides 150 Mbit/s per pipe.

9. David Sincowskie, Bellcore, is working with Batcher-Banyan design and has working 32x32 switches.

10. Stratacom has a commercial product which is really a T1 voice switch implemented internally by a packet switch, where the packet is 192 bits (T1 frame). This switch can pass 10,000 packets per second.

11. Stanford NAB provides 30-50 Mbit/s throughput on 100 Mbit/s connection using Versatile Message Transaction Protocol (VMTP) [see RFC 1045]

12. The December issue of IEEE Journal on Selected Areas in Communications, provides much detail concerning interconnects.

13. Ultranet Technology has a 480 Mbit/s connection using modified ISO TP4.

14. At MIT, Dave Clark has an architecture proposal of interest.

15. At CMU, the work of Eric Cooper is relevant.

16. At Protocol Engines, Inc., Greg Chesson is working on an XTP-based system.

17. Larry Landweber at Wisconsin University is doing relevant work.

18. Honeywell is doing relevant work for NASA.

19. Kung at CMU is working on a system called "Nectar" based on a STARLAN on fiber connecting dissimilar processors.

20. Burroughs (now Unisys) has some relevant work within the IEEE 802.6 committee.

21. Bellcore work in "Switched Multimedia Datanet Service" (SMDS) is relevant (see paper supplied by Dave Clark).

22. FDDI-2, a scheme for making TDMA channel allocations at 200 Mbit/s.

23. NRI, Kahn-Farber Proposal to NSF, is a paper design for high-bandwidth network.

24. Barry Goldstein work, IBM-Yorktown.

25. Bell Labs S-Net, 1280 Mbit/s prototype.

26. Fiber-LAN owned by Bell South and SECOR, a pre-prototype 575 Mbit/s Metro Area Net.

27. Bellcore chip implementation of FASTNET (1.2 Gbit/s).

28. Scientific Computer Systems, San Diego, 1.4 Gbit/s prototype.

29. BBN Monarch Switch, Space Division pre-prototype, chips being fabricated, 64 Mbit/s per path.

30. Proteon, 80 Mbit/s token ring.

31. Toronto University, 150 Mbit/s "tree"--- really a LAN.

32. NSC Hyperchannel II, reputedly available at 250 Mbit/s.

33. Tobagi at Stanford working on EXPRESSNET; not commercially available.

34. Columbia MAGNET-- 150 Mbit/s.

35. Versatile Message Transaction Protocol (VMTP).

36. ST integrated with IP.

37. XTP (Chesson).

38. Stanford Transport Gateway.

39. X.25/X.75.

40. Work of the Internet Activities Board.

B. Gigabit Working Group Members

Member	Affiliation
Gordon Bell	Ardent Computers
Steve Blumenthal	BBN Laboratories
Vint Cerf	Corporation for National Research Initiatives
David Cheriton	Stanford University
David Clark	Massachusetts Institute of Technology
Barry Leiner (Chairman)	Research Institute for Advanced Computer Science
Robert Lyons	Defense Communication Agency
Richard Metzger	Rome Air Development Center
David Mills	University of Delaware
Kevin Mills	National Bureau of Standards
Chris Perry	MITRE
Jon Postel	USC Information Sciences Institute
Nachum Shacham	SRI International
Fouad Tobagi	Stanford University

End Notes

[1] Workshop on Computer Networks, 17-19 February 1987, San Diego, CA.

[2] "A Report to the Congress on Computer Networks to Support Research in the United States: A Study of Critical Problems and Future Options", White House Office of Scientific and Technical Policy (OSTP), November 1987.

[3] We distinguish in the report between development of a backbone network providing gigabit capacity, the GB, and an interconnected set of high-speed networks providing high-bandwidth service to the user, the Gigabit Network (GN).

[4] Incidentally, they already manage to serve 150 million subscribers in an 11-digit address-space (about 1:600 ratio). We have a 9.6-digit address-space and are running into troubles with much less than 100,000 users (less than 1:30,000 ratio).